信息类、导航类与事务类查询的网络动态性分析*

张晓娟
(西南大学计算机与信息科学学院　重庆 400715)

摘要:【目的】分析信息类、导航类与事务类查询随时间的网络动态性特征, 以期为搜索引擎性能优化提供相关依据。【方法】利用相关评测指标分别从查询动态、文档内容动态和信息需求动态三个角度出发, 分析不同意图类别查询随时间变化所呈现的特征; 针对不同意图类别查询, 分析在不同查询流行度特征中, 其文档内容以及信息需求的变化情况。【结果】在查询流行度分布方面, 信息类查询通常包含波峰, 事务类查询更可能包含多个波峰且具有周期性, 导航类查询通常保持平滑趋势; 信息类查询随网页内容与信息需求变化幅度均比其他两类查询的要大。【局限】观察时间段只有 29 天; 未对不包含波峰与包含多个波峰的查询流行度分布图中波峰进行归类与自动识别。【结论】对于信息类查询来说, 搜索引擎尽可能地对其查询结果进行多样化展示; 对于导航类查询来说, 搜索引擎需要保证与之相关权威网页在查询结果中的靠前性; 对于与用户交互行为相关的事务类查询, 应长时间保持相关网页排序不变; 对于一些与娱乐相关事务类查询, 在网页排序中需考虑网页的新颖性。

关键词: 信息类查询　事务类查询　导航类查询　查询动态　信息需求动态　文档内容动态
分类号: G353.4

1 引　言

自 Broder[1]按照用户意图(或用户任务)将查询划分为信息类、导航类、事务类三大类别后, 学界对如何选取分类特征实现三大查询类别之间的有效区分进行了大量研究[2-4]。由于 Broder[1]对查询进行分类的最终目的是为了使搜索引擎能够根据不同意图类别查询为用户提供不同的检索服务, 而获取和分类查询仅仅是手段, 因此如何对归类后的查询进行分析并以此为搜索引擎性能优化提供依据是一重要研究方向。

作为用户交互信息的场所, 网络呈现出动态变化特征。其中, 受 Kulkarni 等[5] 研究的启发, 本文从用户获取信息的角度将网络动态性主要分为查询动态、信息需求动态和文档内容动态三方面。查询动态是指大众用户所提交某查询的频次随时间而发生的变化; 信息需求动态是指用户针对同一查询的大众信息需求随时间而发生的变化, 如用户在马来西亚航空公司 MH370 空难之前提交查询“马航”的一般信息需求是想获得与马航相关航班信息, 而在马航空难发生后, 用户提交该查询的一般信息需求更可能与该空难相关; 文档内容动态性是指在不同时间点与某查询相关文档之间的内容差异性。网络动态性的特征分析有助于搜索引擎从动态角度理解用户意图与网页内容变化规律, 使得检索结果能满足用户即时信息需求。如在查询推荐中, 通过对查询动态性分析, 能为用户推荐当前较为流行的查询; 在查询结果排序中, 通过对信息需求动态和文档动态分析, 能准确定位与用户最新信息需求相关的文档。因此, 如何使检索结果适应网络动态性特征是搜索引擎性能优化中需考虑的重要方面之一。

在信息检索中, 用户意图类别被认为是很重要的

通讯作者: 张晓娟, ORCID: 0000-0002-5889-5922, E-mail: zhangxiaojuan624@gmail.com。

*本文系国家社科基金青年项目“融合用户个性化与实时性意图的查询推荐模型研究”(项目编号: 15CTQ019)和西南大学博士启动基金“查询意图自动分类与分析研究”(项目编号: SWU114093)的研究成果之一。

用户情景因素，直接影响用户获取信息的途径以及想要获取信息的类型[6]。因此，搜索引擎在为适应网络动态性而进行性能优化时，也需考虑相关的用户意图因素。鉴于此，本文将对不同任务类别查询(如信息类、导航类与事务类查询)的网络动态性进行比较与分析，以期为搜索引擎针对不同用户意图的网络动态性能优化提供相关依据。

2 国内外研究现状

2.1 查询意图相关研究

2002 年，Broder[1]通过用户调研与对 AltaVista 查询日志分析将查询意图分为信息类、导航类和事务类。信息类意图(如查询“竞价广告”，“如何减肥”)指用户以一种静态方式去查询被认为能在网络上获取到的信息，除阅读之外无其他交互信息，查找内容可以是数据、文档、文或多媒体，信息需求既可以是精确的又可以是模糊的；导航类(如查询“国家留学基金委网站”，“阿姆斯特丹大学 主页”)指用户查找某个特定网站(网页)，该网站(网页)可以是个人网站(网页)也可以是组织网站(网页)等，即用户在执行检索时已在头脑中形成了查找意向，知道或者认为存在网址可以满足自己的信息需求；事务类意图(如查询“七里香 下载”，“Gmail 注册”)指用户通过查找获取一些资源或网络服务，比如购买、下载等。在文献[1]的基础上，文献[2-4]探讨了如何选取分类特征，以此实现这三类查询的自动区分；另外还有学者尝试通过对不同意图类别查询的特征进行分析，以此为不同意图类别查询构建检索模型，如 Fujii[6]首先根据查询词在网页锚文本中分布情况识别该查询是事务类还是信息类查询，且通过分析发现导航类查询适合基于锚文本的检索方法，而信息类查询适合基于内容的检索方法；Craswell 等[7]为信息类与导航类查询提出了不同的检索排序模型，其实验结果表明，基于链接排序的方法能够有效提高导航类查询的检索性能；Ali 等[8]通过对信息类、导航类与事务类查询分别在搜索引擎 Yahoo 和 Google 中检索结果对比分析得知，Google 针对事务类查询的检索准确率最高，而 Yahoo 对信息类查询的检索准确度最高。

2.2 网络动态性相关研究

如 Kulkarni 等[5]的归类，网络动态性研究主要包括查询动态性、信息需求动态性和文档内容动态性三方面，本文也从这三方面对其加以综述。

(1) 查询动态性研究主要集中在通过观察查询随时间变化规律预测一些社会现象，如 Beitze 等[9]利用查询日志数据分析每小时内查询流行度(Query Popularity)与查询主题的变化情况；Valchos 等[10]首次尝试利用傅里叶分析为网络查询的周期性和突发性建模；Ginsberg 等[11]通过分析大量查询在查询日志中出现情况跟踪流感疾病在人群中的爆发情况；Adar 等[12]利用查询词频的变化理解用户以往行为并推测其将要发生的行为。

(2) 信息需求动态性研究主要集中在如何构建模型定位用户实时性意图从而实现查询推荐或者检索结果重排序，如 Johansson 等[13]通过构建图模型表征查询与潜在用户意图的动态关系，以此能在不同时间点获得与原始查询信息需求相关的候选查询推荐；Whiting 等[14]根据查询在查询日志中出现频率捕捉用户最新信息需求，以此实现能满足用户实时性意图的查询推荐；Alonso 等[15]通过利用文档的时间片段建立检索模型；Berberich 等[16]利用数学建模为不同时间段查询提供不同的多样化检索结果。

(3) 文档内容动态性的主要相关研究集中在如何采用相关方法衡量网页文档内容变化规律，如 Cho 等[17]基于词级别对 4 个月内网页内容的变化情况进行分析发现，约 40%的网页内容每周都会发生变化；Fetterly 等[18]首先分析每个网页内容随时间的变化程度，再分析与变化程度相关的因素，其实验发现网页内容变化程度与该网页的域名相关；Ntoulas 等[19]利用词级别分析网页内容变化情况；Kim 等[20]利用网页相关用户行为(如下载频率、修改频率等)衡量网页随时间变化情况；Cho 等[21]利用网页中的超链接信息衡量网页变化情况；Adar 等[22]基于网页中 DOM 元素以及单个词随时间变化情况，提出衡量网页内容变化的算法和模型；Kausar 等[23]提出根据网页中哈希码(Hash)的变化情况判断网页内容是否发生了变化。总之，基于词信息是衡量网页内容变化的最常用方法。

另外，还有学者探讨了如何利用网络动态特征(如查询动态与网页内容动态)提高搜索引擎性能，如 Alonso 等[24]提出一种基于文本中时间表达式对查询结果进行聚类的方法；Alfonseca 等[25]研究表明查询周期性能提高查询建议的准确度；Dakka 等[26]开发了能

针对同一问题在不同时间点提供不同答案的问答系统；Zahedi 等[27]首先识别查询中包含的时间段信息，再将其融合博客检索模型中，以此返回特定时间段的博客信息；Elsas 等[28]将时间属性融合到语言模型中，以此提高导航类查询检索结果的准确度；Syed 等[29]提出一种能根据查询中不同用户意图而提供不同检索结果的检索模型。综合已有研究可以看出，目前仍未有对信息类、导航类与事务类查询进行网络动态性分析的相关研究。

3　衡量网络动态性的方法

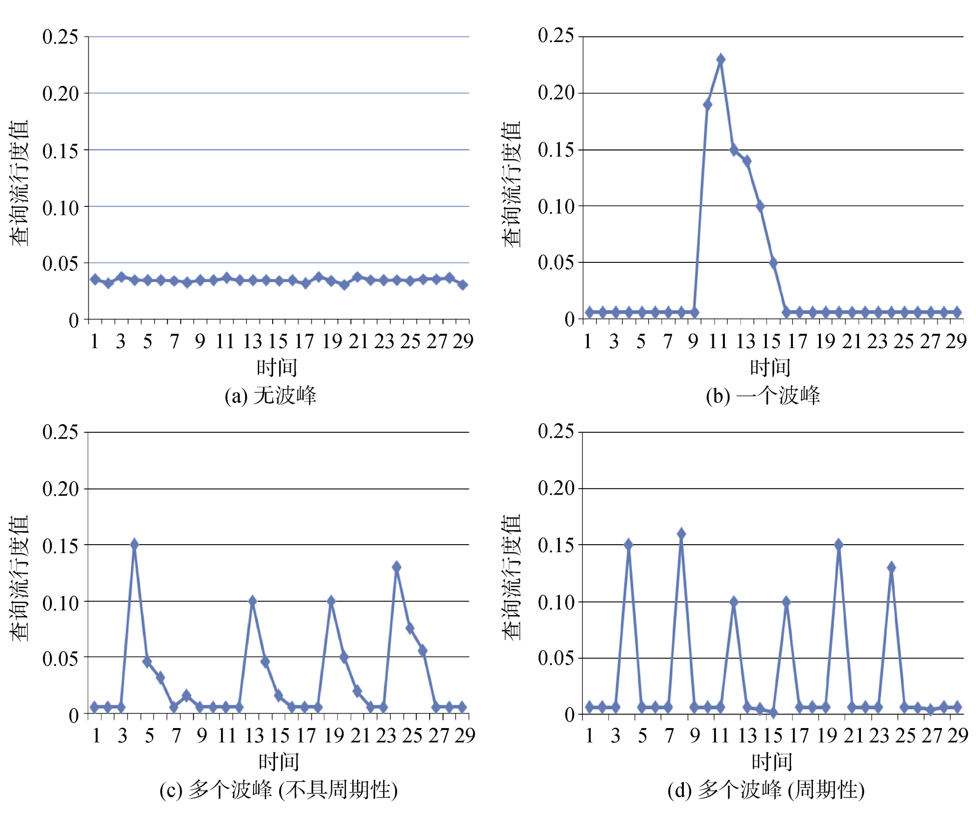
3.1　衡量查询动态的方法

本文采用查询流行度分布[30]衡量查询动态性。其中，查询流行度分布是指在特定时间范围内，查询在具体每天出现频次与该时间范围内总出现频次比值的分布。为了能从更深层次揭示查询动态特征，本文根据 Kulkarni 等[5]从波峰数、波峰形状与整体趋势对查询流行度分布的归类，提出自动识别其不同类别的相关方法。

(1) 基于波峰数的查询流行度分布分类

①波峰类别

从查询流行度分布中所包含的波峰数角度，查询流行度分布可分为不包含波峰、包含一个波峰以及包含多个波峰三类别，如图 1 所示。当查询流行度分布中包含多个波峰时，又可将其细分为具有周期性与不具有周期性两类，如图 1(c)–图 1(d)所示。其中，图 1 中横轴表示具体某一天(本文中观察时间范围为 29 天)，纵轴表示查询流行度值 。



(a) 无波峰　(b) 一个波峰　(c) 多个波峰 (不具周期性)　(d) 多个波峰 (周期性)

图 1　查询流行度分布中的波峰类别

②波峰类别识别

从图 1 中可以看出，当查询流行度分布中存在波峰时，该波峰对应的查询概率值远远大于其邻近时间点的概率值，本文将此概率值所对应的时间点称为突发点。据观察可知，每个波峰对应一个突发点。因此，本文通过对突发点的识别判断某查询流行度分布中的波峰数，且所采用方法是移动平均方法[10]。当查询流行度分布中存在多个突发点时，若每两个突发点之间的时间间隔相等则说明该查询在观察时间内的查询流行度随时间分布具有周期性；反之，不具周期性。

(2) 基于波峰形状的查询流行度分布分类

①查询流行度分布的波峰形状

当某查询流行度分布中包含一个波峰时，其波峰形状可分为以下 4 类：

1) Wedge(楔子)：即查询流行度分布中出现波峰前后时

间点的查询流行度随时间上升和下降的速率相同，如图 2(a)所示；

2) Castle (城堡)：即查询流行度分布中出现波峰后，其查询流行度值在后续一段时间内保持稳定，如图 2(b)所示；

3) Sail(帆状)：在某时间内迅速上升到峰值后再缓慢地下降，如图 2(c)所示的右帆状；或查询流行度在某段时间内缓慢上升到波峰值后再在较短时间内迅速下降(1 或 2 天之内)，如图 2(d)所示的左帆状。

需指出的是，当查询中不包含波峰值或者当查询中包含多个波峰值时，其整体形状的划分不在本文探讨范围之内。

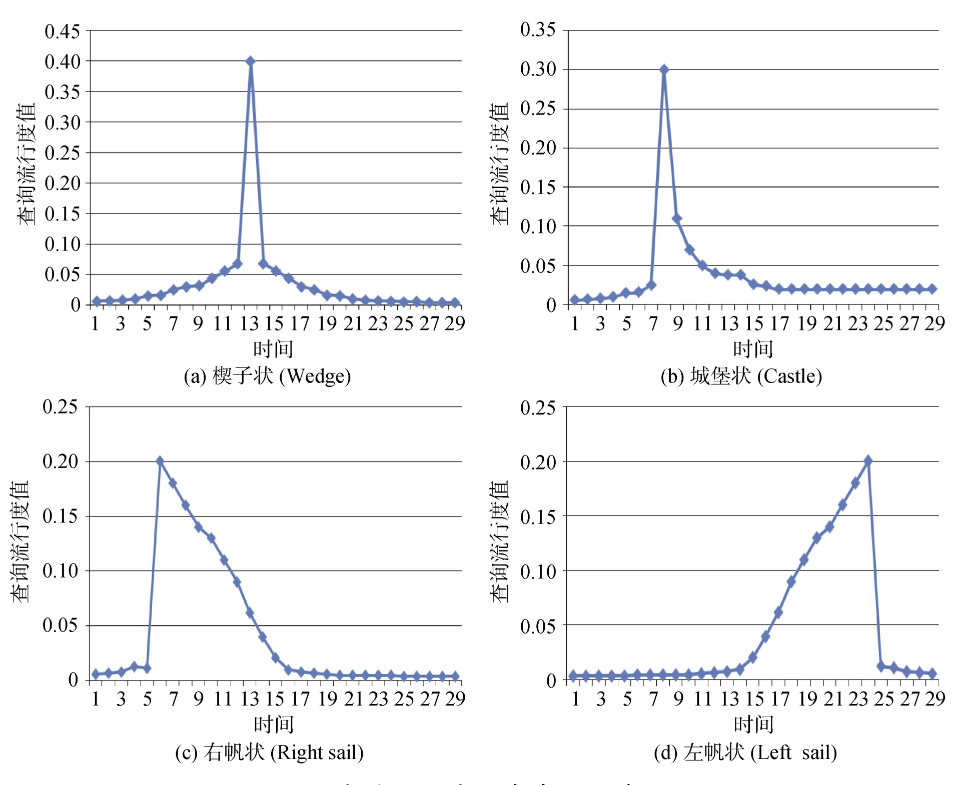


图 2　查询流行度分布中的波峰形状

②查询流行度整体形状识别

获得查询流行度分布中最大概率值 P_t 以及相应的时间点 t，若该时间点前的相邻两时间点概率之差 $P_{t-s}-P_{t-s-1}$ ($2\leqslant s\leqslant n-t-1$，$n$ 为总观察天数)与该时间点后的相邻两时间点之差 $P_{t+s}-P_{t+s+1}$ 的差值绝对值(即 $|(P_{t+s}-P_{t+s+1})-(P_{t-s}-P_{t-s-1})|$)小于一定阈值(本文设定为 0.0005)，则该波峰形状为楔子形状。

当查询流行度分布不是楔子形状时：若在时间点 t 后，对于任意一点 m ($t+1\leqslant m\leqslant n$，$n$ 表示总观测天数)来说，若 $|P_m-P_{m-1}|$ 与 $|P_{m+1}-P_m|$ 之间差值绝对值小于一定阈值(本文设定为 0.0005)，则该波峰呈城堡形状；若 P_t-P_{t-1} 值大于一定阈值(本文设定为 $8\times P_{t-1}$)，而 P_t-P_{t+1} 值小于一定阈值(即 $0.01\times P_{t+1}$)，则该波峰呈现左帆状；若 P_t-P_{t-1} 值小于一定阈值(本文将其设定为：$0.01\times P_{t-1}$)，而 P_t-P_{t+1} 值大于一定阈值(即 $8\times P_{t+1}$)，则该波峰呈右帆状。

(3) 基于整体趋势的查询流行度分布分类

图 3 分别表示查询流行度整体趋势所属的类别，即向上趋势、向下趋势、平滑趋势以及上升–下降趋势。在观察时间内，若不存在突发点，则该趋势为平滑；若约有不少于 75%的概率值(P_t)大于其后一时间段的概率值 P_{t-1} ($2\leqslant t\leqslant n$，n 为总观察天数)，其整体趋势呈向上趋势；若有不少 75%的概率值 P_t 小于其后一时间段的概率值 P_{t-1} ($1\leqslant t\leqslant n$，n 为总观察天数)，其整体趋势呈向下趋势；若存在一个突发点，其整体趋势呈上升–下降趋势。

3.2　衡量信息需求动态的方法

点击信息是表征用户信息需求的重要来源，用户针对同一查询的点击信息变化也能在一定程度上体现用户针对该查询的信息需求变化情况[30]。基于此，本文基于点击熵(Click Entropy)[31]衡量用户信息需求变化情况，其计算方法如公式(1)所示。

$$clickEntropy(q)=\sum_{d\in D(q)}-P(d\mid q)\log_2 P(d\mid q)\quad(1)$$

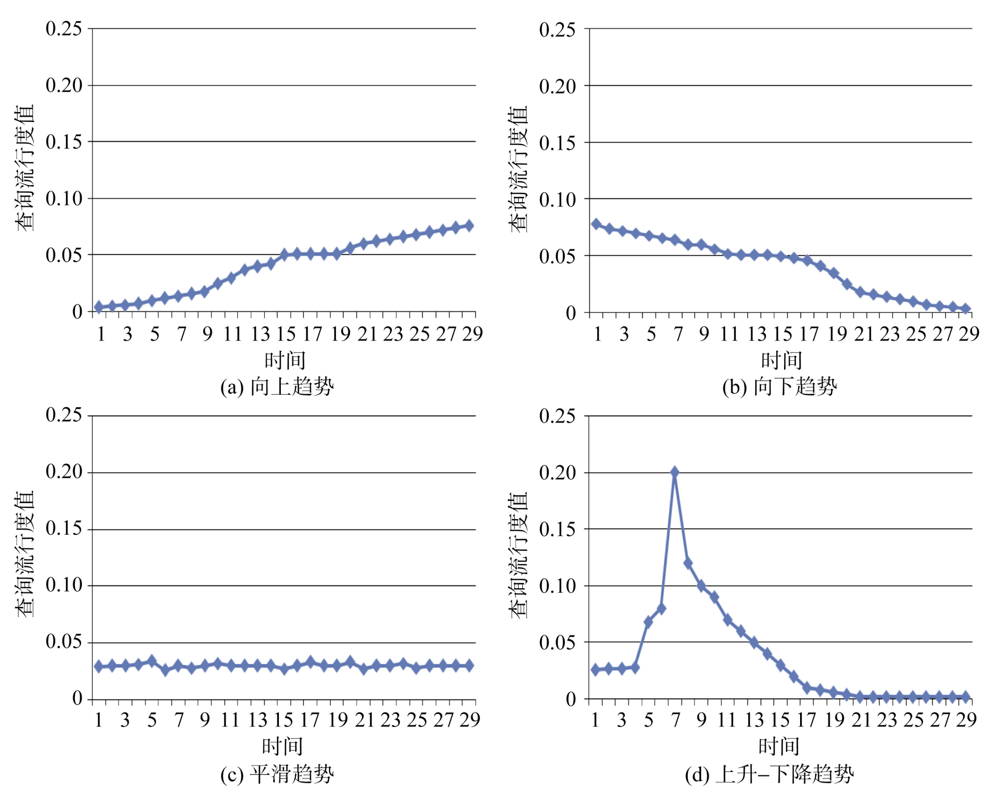


图 3 查询流行度分布中整体趋势图

其中，$clickEntropy(q)$ 表示查询 q 的点击熵值，D 表示用户提交查询 q 后点击的文档集合，$P(d|q)$ 表示提交查询 q 后点击文档 d 的概率。本文假设若用户针对某查询在不同时间点的点击熵值变化越大，表明用户信息需求随时间变化越大；反之，用户信息需求随时间变化越小。为了获得某查询随时间的信息需求变化情况，首先计算观察时间段(本文设定为 29 天)内每个观察时间点 t（$2 \leqslant t \leqslant 29$）与前一个观察时间点 $t-1$ 之间点击熵的差值绝对值，再对观察时间内所有点击熵的差值绝对值求平均值(记作 *AvgClickEntropy*)得到该查询的信息需求变化值。其中，*AvgClickEntropy* 值越大表明查询随时间的信息需求变化越大，反之越小。需说明的是，本文在此探讨是大众用户的信息需求，而非个性化需求。

3.3 衡量文档动态方法

本文采用基于文档中词信息来衡量网页内容变化情况，采用指标主要基于单个词变化的 TF-IDF 余弦距离值[5]和基于词串变化的 Shingle 距离[31]。

(1) TF-IDF 余弦距离值

$D_{\cos}(P_1, P_2)$ 用于计算给定两个不同网页形式 P_1 与 P_2 之间的 TF-IDF 余弦距离值，如公式(2)所示。

$$D_{\cos}(P_1, P_2) = 1 - \frac{v_1 v_2}{\|v_2\|_2 \|v_1\|_2} \quad (2)$$

其中，v_1 与 v_2 分别表示网页 P_1 与 P_2 的 TF-IDF 权值向量。$v_1 v_2$ 表示向量 v_1 与 v_2 的内积，$\|v_1\|_2$ 与 $\|v_2\|_2$ 分别表示向量 v_1 与 v_2 的模。$D_{\cos}(P_1, P_2)$ 值越大，说明两网页之间差异性越大；反之，其值越小。

(2) Shingle 距离

Shingle 距离[31]首先将文档中连续几个字符串视为该文档的一个 Shingle，如 将某文档连 4 个字符串组成一个 Shingle，则该文档中的一段文本(a, rose, is, a, rose, is, a, rose)的 Shingle 集合为：{(a, rose, is, a), (rose, is, a, rose), (is, a, rose, is)}。再利用公式(3)衡量两不同网页之间 Shingle 的变化情况。

$$ShSim(D) = \frac{|Sh(D^1) \cap Sh(D^2)|}{N} \quad (3)$$

其中，$Sh(D^1)$ 与 $Sh(D^2)$ 分别表示文档 D^1 与 D^2 中包含的 Shingles 集合，N 表示两个文档中 Shingle 集合个数。衡量文档随时间变化的指标 $ShDiff(D) =$

$1-ShSim(D)$，该指标值越大，表明网页之间内容差距越大。本文将文档中连续三个字符串(非停用词)视为一个 Shingle。

为获得某查询 q 在观察时间段(本文设为 29 天)内的文档内容变化值(记作 *ContentChange*)，笔者首先为该查询在每个观察时间点获得点击排序前 5 的文档集合，再分别利用TF-IDF余弦和Shingle距离计算每个观察点 $t(2\leqslant t\leqslant 29)$的文档集合与前一个观察点 t–1 的文档集合之间任意两文档之间内容差异值，经平均值后可得到该查询在某时间相对前一时间的内容变化值(记作 $m_t(q)$)，再对该观察时间段内所有 $m_t(q)$ 值求平均从而可获得该查询的 *ContentChange* 值。该值越大，说明查询随时间的文档内容变化越大；反之，越小。

4 实验数据集

笔者采用 Sogou 实验室发布的 2008 年 6 月(6 月 1 日–6 月 29 日)查询日志[①]数据作为实验数据集，数据格式如表 1 所示。

表 1 Sogou 查询日志数据格式样

用户访问时间	用户 ID	查询词	用户点击 URL 在返回结果中的排名	用户点击的顺序号	用户点击的 URL
00:00:03	35804326352621896	[免费取名]	3	1	http://huaxia.wangzhan8.com/
00:00:03	07321773511158924	[欧美金发女郎]	2	4	http://a.se2222.com/Html/OPIC/index.html
00:00:03	43080219994871455	[google]	1	1	http://www.google.com/

受人力以及时间的限制，本文无法对查询日志数据中的每个查询进行分析，故笔者首先利用泊松抽样方法[31]从 Sogou 日志数据集中抽取了 3 000 个查询，且这些查询满足以下条件：在查询日志中出现频次不少于 2 000；每天至少包含 5 个不同的被点击 URL。

笔者要求三位标注者对这 3 000 个查询应属类别(信息类、导航类与事务类)进行人工标注。考虑到一个查询可能属于多种类别，例如，给定查询“MP3”，用户潜在意图可能有：了解 MP3 的相关信息，即信息类查询；到达某个 MP3 网站，即导航类查询；下载 MP3 格式的文件，即为事务类查询。因此，笔者在此要求标注者标记出查询在大多数情况下应该属于的类别。在标注过程中，若某查询的用户意图类别难以判断时，将由三位标注者共同商讨决定。经过人工标注，分别获得 305 个导航类查询，752 个事务类查询和 1 943 个信息类查询。最后再通过随机抽样方法分别从每个类别查询中各抽取 100 个样本查询用于实验分析。为探讨与查询相关网页内容随时间变化情况，笔者为每个查询每天选取点击频次排名前 5 的 URL，再利用爬虫程序抓取了每个 URL 的网页内容，最后再对每个网页进行正文提取[②]、分词处理[③]等处理。

5 实验结果分析

5.1 查询动态分析

在给定 Sogou 查询日志数据集中，笔者对所选取信息类、导航类与事务类查询的查询动态进行统计分析，得到各意图类别查询在不同查询流行度分布中的比值情况，如表 2 所示。

表 2 信息类、导航类与事务类查询在各类查询流行度分布中的比值

波峰特征 \ 查询类别		信息类	导航类	事务类
波峰数	无波峰	32%	90%	36%
	一个波峰	59%	10%	36%
	多个波峰	9%	0%	28%
周期性	No	8%	0	18%
	Yes	1%	0	10%
波峰形状	城堡	2%	0%	1%
	左帆状	6%	0%	7%
	右帆状	38%	8%	3%
	楔子	13%	0%	28%
整体趋势	向下	25%	0%	22%
	平滑	10%	68%	23%
	向上	20%	17%	45%
	上升–下降	45%	15%	10%

①http://www.sogou.com/labs/dl/q.html.
②https://github.com/stanzhai/Html2Article.
③https://github.com/NLPIR-team/NLPIR.

可知针对查询流行度分布中所包含波峰数来说，大多数(90%)导航类查询不包含波峰，说明用户对导航类查询相关主题的需求比较稳定。据笔者观察可知，不包含波峰的导航类查询大多数(约76%)与一些事业单位相关，而包含波峰的导航类查询大多数(约71%)与公司名相关。超过一半(68%)的信息类查询包含波峰，且这些查询大多(81%)与新闻事件相关，而不包含波峰的信息类查询多与某些概念相关(如查询"搜索引擎 原理")；事务类查询约64%包含波峰，且这些查询大多与电视节目相关；当查询流行度中包含多个波峰时，信息类与事务类查询中查询流行度分布都更有可能不具周期性。通过数据对比可知，信息类查询中查询流行分布更可能包含一个波峰，而事务类查询中查询流行度分布更可能包含多个波峰且更有可能具有周期性。

对于查询中只包含一个波峰时的波峰形状来说，信息类(约38%)与导航类查询(约8%)更有可能呈现右帆状，说明用户对信息类与导航类查询的兴趣更可能是在短时间内产生，且在后续时间内兴趣度是逐渐下降；而事务类查询更有可能(约28%)呈现楔子状，说明了用户对事务类相关主题的兴趣产生和消失速度一致。

对于查询查询流行度中总体趋势来说，信息类查询与事务类查询更有可能(约45%)呈现上升–下降趋势的，说明用户在特定时间类对信息类查询的兴趣更可能具有波动性；导航类查询更有可能(约68%)呈现平滑趋势，说明用户对此类查询的信息需求更可能具有稳定性，且据观察可知，呈平滑趋势的导航类查询多与某组织机构相关(如查询"北京大学")，而呈向上趋势与上升–下降趋势的导航类查询多与公司名相关或者某名人主页相关(如查询"刘德华 博客")；对于事务类查询来说，也更有可能呈现向上趋势，说明用户对事务类查询相关主题的关注度更可能随时间上升。另笔者发现，呈现平滑趋势的事务类查询大多与用户交互行为相关，比如说查询"yahoo邮箱注册"，而呈现向上、上升–下降趋势的事务类多与娱乐活动信息相关，如游戏下载或者电视节目观看等。数据对比结果可以说明，导航类查询保持平滑趋势的概率更高，信息类查询呈现上升–下降趋势的概率更高，而事务类查询呈现向上趋势的概率更高。

5.2　信息需求动态性分析

基于所选取的查询日志数据集，笔者分别计算出信息类、导航类与事务类查询在观察内的 *AvgClickEntropy* 值，其结果如表 3 所示 。

表 3　信息类、导航类与事务类查询的 *AvgClickEntropy* 值

查询类别	*AvgClickEntropy* 值
信息类	3.31
导航类	1.78
事务类	1.17

从表 3 数据可知，信息类查询随时间的信息需求变化幅度大于其他两类查询。为了探讨不同意图类别查询 *AvgClickEntropy* 值的差异性，笔者采用两独立样本 t 检验进行分析，其结果如表 4 所示。

表 4　信息类、导航类与事务类查询间信息需求变化差异度

查询类别	t 统计量的观测值
信息类与导航类	32.64*
导航类与事务类	1.04
信息类与事务类	21.21*

(注: * 表示显著性水平: $p<0.05$)

从表 4 数据可知，信息类与其他两类意图类别查询之间存在着显著性差异(信息类与导航类之间：t 统计量的观测值为 32.64，置信度概率 $p<0.05$；信息类与事务类之间：t 统计量的观测值为 21.21，置信度概率 $p<0.05$)。

5.3　文档内容动态性分析

基于为样本查询采集到的结果集数据，笔者分利用 TF-IDF 与 Shingle 两指标计算不同类别查询中的 *ContentChange(q)*平均值，最终结果如表 5 所示。

表 5　信息类、导航类与事务类查询中的 *ContentChange*(q) 平均值

查询类别	TF-IDF 平均值	ShDiff 平均值
信息类	0.46	0.34
导航类	0.23	0.19
事务类	0.32	0.25

从表 5 数据可知，相对事务类和导航类查询，信息类查询随时间的网页内容变化较大，而导航类查询的网页内容变化幅度最小。为了探讨不同意图类别查询间 TF-IDF 平均值值与 *ShDiff* 平均值之间的差异性，笔者使用两独立样本 t 检验进行分析，其结果如表 6 所示。

表 6　信息类、导航类与事务类查询之间随时间的网页内容变化差异度

查询类别	TF-IDF 平均值	ShDiff 平均值
信息类与导航类	23.10*	13.40*
导航类与事务类	0.25*	0.44*
信息类与事务类	2.45*	5.23*

(注: * 表示显著性水平: $p<0.05$)

从表 6 数据可知, 信息类与其他两类意图类别查询之间存在着显著性差异(信息类与导航类之间 TF-IDF 平均值: t 统计量的观测值为 23.10, 置信度概率 p<0.05; 信息类与事务类之间 TF-IDF 平均值: t 统计量的观测值为 2.45, p<0.05; 信息类与导航类之间 *ShDiff* 平均值: t 统计量的观测值为 13.40, 置信度概率 p<0.05; 信息类与事务类之间平均值: t 统计量的观测值为 5.23, p<0.05。)

5.4　信息需求动态与文档动态随查询动态的分析

相对其他两类网络动态性特征来说, 查询流行度特征更容易观察到。因此, 笔者尝试探讨不同类别查询在不同查询流行度中其信息需求动态和网络动态的特征, 以期能通过较易观察特征推测其隐含特征提供相关依据。

(1) 信息需求动态分析

为了获得信息类、导航类与事务类查询在不同查询动态中信息需求变化情况, 分别计算不同意图类别查询在不同查询动态特征中 *AvgClickEntropy* 值, 其结果如表 7 所示。

表 7　信息类、导航类与事务类查询在不同查询动态中 *AvgClickEntropy* 平均值

波峰特征 \ 查询类别		信息类	导航类	事务类
波峰数	无	0.02	0.11	0.23
	一个波峰	1.74	0.81	1.01
	多个波峰	3.52	–	2.34
周期性	Yes	5.51	–	3.28
	No	3.52	3.24	2.34
波峰形状	城堡	0.09	1.54	0.09
	左帆状	1.52	–	1.52
	右帆状	1.52	1.48	1.50
	楔子	3.12	–	2.24
整体趋势	下降	4.45	–	4.35
	上升	2.53	1.70	2.31
	平滑	1.12	0.71	1.13
	上升–下降	5.24	2.09	4.03

从表 7 数据可知, 只包含一个波峰查询的 *AvgClickEntropy* 值低于包含多个波峰的查询的 *AvgClickEntropy* 值, 说明包含多个波峰的查询中包含的用户信息需求随时间变化幅度越大; 当不同意图类别查询包含相同波峰时, 信息类查询中信息需求变化幅度大于其他两类查询。无论查询流行度分布是否具有周期性来, 信息类查询的信息需求变化幅度相对其他两类查询要大。

对于查询流行度分布中的波峰形状来说, 当波峰形状呈楔子时, 信息类查询的信息需求变化幅度大于事务类查询; 当波峰形状呈城堡形状时, 事务类查询的信息需求变化幅度大于信息类查询; 当波峰形状呈帆状时, 信息类、导航类和事务类查询的信息需求变化幅度几乎一致。对于查询流行度的不同整体趋势来说, 分布趋势为平滑查询的信息需求变化幅度相对较小, 而分布趋势为上升或者下降趋势查询的信息需求变化幅度相对较大。且针对不同整体趋势来说, 信息类查询的信息需求变化幅度相对其他两类查询要大。

(2) 文档内容动态分析

为获得信息类、导航类与事务类查询在不同查询动态特征中的网页内容变化情况, 笔者分别计算了不同意图类别查询在各查询动态特征中相应的 *ContentChange(q)* 平均值, 具体结果如表 8 所示。

表 8　信息类、导航类与事务类查询在不同查询流行度特征中的网页内容变化情况

查询流行度类别		ContentChange(q) (TF-IDF) 信息类	导航类	事务类	ContentChange(q) (ShDiff) 信息类	导航类	事务类
波峰数	无波峰	0.10	0.09	0.20	0.41	0.18	0.35
	一个波峰	0.42	0.19	0.30	0.44	0.32	0.43
	多个波峰	0.49	–	0.41	0.52	–	0.44
周期性	Yes	0.44	0.32	0.34	0.43	0.20	0.27
	No	0.49	–	0.45	0.57	–	0.38
波峰形状	城堡	0.30	0.21	0.41	0.43	0.42	0.33
	左帆状	0.38	–	0.42	0.35	–	0.40
	右帆状	0.36	0.38	0.38	0.34	0.35	0.38
	楔子	0.52	–	0.54	0.48	–	0.52
整体趋势	平滑	0.54	0.45	0.52	0.61	0.41	0.57
	下降	0.52	–	0.52	0.52	–	0.52
	上升	0.32	0.27	0.31	0.42	0.30	0.42
	上升–下降	0.20	0.19	0.21	0.29	0.19	0.28

可知，在观察时间内，查询流行度分布中包含的波峰越多，与该查询相关的文档内容变化幅度越大。当不同意图类别查询包含相同波峰时，信息类查询的网页内容变化幅度大于其他两类查询。在查询流行度分布的周期性中，周期性查询的网页内容变化幅度小于非周期查询的变化幅度；当查询流行度分布具有周期性时，信息类查询的网页内容变化幅度大于事务类查询的网页内容变化幅度。周期性的事务类查询通常与当前流行的电视节目或者体育事件相关，周期性的信息类查询大多与名人相关。

在查询流行度的波峰形状中，楔子形状的网页内容变化幅度分别大于帆形状与城堡形状的网页内容变化幅度。针对同一波峰形状，事务类查询的网页内容变化幅度大于其他两类查询的网页内容变化幅度。在查询流行度的整体趋势中，呈现上升–下降趋势查询的网页内容变化的幅度较少，而平滑与下降趋势的查询的网页内容变化幅度较大，其主要原因在于，呈现上升与下降趋势的查询包含不同查询分面，用户在不同时刻对不同分面感兴趣，故在不同时刻查询其相关的文档内容存在着差异性。另外，针对同一整体趋势，导航类查询的网页内容变化幅度小于其他两类查询，说明用户更偏好搜索引擎能为导航类查询在不同时间段返回内容比较一致的网页。

6 搜索引擎性能优化的相关建议

根据以上实验结果分析，笔者对搜索引擎相关性能优化提出了以下建议。

(1) 对于信息类查询来说，需随时捕捉查询中可能包含的潜在用户意图，且尽可能地对其查询结果进行多样化，能为查询的不同分面返回相关信息，满足用户多样化需求；另对查询流行度包含波峰的信息类查询，在波峰产生后的短时间内(3 天–5 天)，可优先做为相关查询的候选查询推荐。

(2) 对于导航类查询来说，用户的信息需求具有明确性，则搜索引擎需要保证与之相关权威网页在查询结果中的靠前性；另导航类查询包含的用户信息需求比较固定，则搜索引擎针对此类查询返回的网页内容可保持不变，且可利用长时间信息(如用户过去行为信息)优化查询结果。

(3) 对于与用户交互行为相关的事务类查询，其用户需求和网页内容随时间变化幅度小，故在长时间保持相关网页排序不变；对于一些与娱乐相关事务类查询，用户可能周期性对最新事件感兴趣，故搜索引擎可周期性抓取最新网页，并将其融合到查询结果中，且在网页排序中需考虑网页的新颖性。

7 结 语

本文主要从查询动态、信息需求动态和文档内容动态三方面对信息类、导航类与事务类查询的网络动态性进行分析。另外，还进一步分析了不同意图类别查询信息需求动态与文档内容动态随查询动态的情况。 最后，还对搜索引擎性能优化提出了相关建议。尽管如此，本文还存在一些不足之处，也是笔者后续研究工作中还需进一步探讨的内容：从更长时间范围探讨网络动态变化特征；进一步对不包含波峰与包含多个波峰的查询流行度的波峰进行归类且提出自动识别的方法；综合考虑文档结构及其词变化来识别文档内容变化情况。

参考文献：

[1] Broder A. A Taxonomy of Web Search[J]. SIGIR Forum, 2002, 36(2) : 3-10.

[2] 伍大勇，赵世奇，刘挺，等. 融合多类特征的 Web 查询意图识别[J]. 模式识别与人工智能，2012，25(3): 500-505. (Wu Dayong, Zhao Shiqi, Liu Ting, et al. Identification of Query Intent via Combining Multiple Features[J]. Pattern Recognition and Artificial Intelligence, 2012,25(3): 500-505).

[3] Figueroa A. Exploring Effective Features for Recognizing the User Intent Behind Web Queries[J]. Computers in Industry, 2015, 68: 162-169.

[4] Zamora J, Mendoza M, Allende E. Query Intent Detection Based on Query Log Mining[J]. Journal of Web Engineering, 2014, 13(1): 24-52.

[5] Kulkarni A, Teevan J, Svore K M, et al. Understanding Temporal Query Dynamics[C]// Proceedings of the 4th International Conference on Web Search and Web Data Mining, Hong Kong, China. 2011.

[6] Fujii A. Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval[C]//Proceedings of the 17th International Conference on World Wide Web. 2008.

[7] Craswell N, Hawking D, Robertson S. Effective Site Finding Using Link Anchor Information[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research

and Development in Information Retrieval. 2001: 250-257.

[8] Ali S, Gul S, Gorman, G E. Search Engine Effectiveness Using Query Classification: A Study[J]. Online Information Review, 2016, 4(40): 515-528.

[9] Beitzel S M, Jensen E C, Chowdhury A, et al. Hourly Analysis of a Very Large Topically Categorized Web Query Log[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004: 321-328.

[10] Vlachos M, Meek C, Vagena Z. Identifying Similarities, Periodicities and Bursts for Online Search Queries[C]// Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. 2004:131-142.

[11] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting Influenza Epidemics Using Search Engine Query Data[J]. Nature, 2009, 457(7232): 1012-1014. DOI: 10.1038/nature 07634.

[12] Adar E, Weld D, Bershad B, et al. Why We Search: Visualizing and Predicting User Behavior[C]// Proceedings of the 16th International Conference on World Wide Web. 2007: 161-170.

[13] Johansson F, Färdig T, Jethava V, et al. Intent-aware Temporal Query Modeling for Keyword Suggestion[C]// Proceedings of the 21st ACM International Confenrence on Information and Knowledge Managent. 2012: 83-86.

[14] Whilting S, McMinn A J, Jose J M. Exploring Real-Time Temporal Query Auto-Completion[C]// Proceedings of the 13th Dutch-Belgain Workshop on Information Retrieval. 2013: 12-15.

[15] Alonso O, Baeza-Yates R, Gertz G. Effectiveness of Temporal Snippets[C]//Proceedings of the 18th International Conference on World Wide Web. 2009.

[16] Berberich K, Bedathur S. Temporal Diversification of Search Results[C]// Proceedings of the SIGIR 2013 Workshop on Time-aware Information Access. 2013.

[17] Cho J, Garcia-Molina H. The Evolution of the Web and Implications for an Incremental Crawler [C]// Proceedings of the 26th International Conference on Very Large Databases. 2000.

[18] Fetterly D, Manasse M, Najork M, et al. A Large-scale Study of the Evolution of Web pages[C]// Proceedings of the 18th International Conference on World Wide Web. 2003.

[19] Ntoulas A, Cho J, Olston C. What's New on the Web? The Evolution of the Web from a Search Engine Perspective[C]// Proceedings of the 13th International Conference on World Wide Web. 2004.

[20] Kim S J, Lee S H. An Empirical Study on the Change of Web Pages[A]// Web Technologies Research and Development [M]. Springer Berlin Heidelberg, 2004: 632-642.

[21] Cho J, Roy S, Adams R E. Page Quality: In Search of an Unbiased Web Ranking[C]// Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. 2005: 551-562.

[22] Adar E, Teevan J, Dumais S T, et al. The Web Changes Everything: Understanding the Dynamics of Web Content [C]// Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. 2009.

[23] Kausar A, Dhaka V S, Singh S K. A Novel Web Page Change Detection Approach Using SQL Server[J]. Journal of Modern Educatin and Computer Science, 2015, 9(7): 36-43.

[24] Alonso O, Gertz M. Clustering of Search Results Using Temporal Attributes[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA. 2006.

[25] Alfonseca E, Ciaramita M, Hall H, et al. Lexical Relationships from Temporal Patterns of Web Search Queries[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009.

[26] Dakka W, Gravano L, Ipeirotis P G. Answering General Time Sensitive Queries[C]//Proceedings of the ACM 17th Conference on Information and Knowledge Management. 2008.

[27] Zahedi M, Aleahmad A, Rahgozar M, et al. Time Sensitive Blog Retrieval Using Temporal Properties of Queries[J]. Journal of Information Science, 2015, 43(1): 1-19. DOI: 10.1177/0165551515618589.

[28] Elsas J, Dumais S T. Leveraging Temporal Dynamics of Document Content in Relevance Ranking[C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. 2010: 1-10.

[29] Syed U, Slivkins A, Mishra N. Adapting to the Shifting Intent of Search Queries [C]// Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. 2009.

[30] Broder A Z, Glassman S C, Manasse M S. Syntactic Clustering of the Web[J]. Journal of Computer Networks and ISDN Systems, 1997,29(8-13): 1157-1166.

[31] Ozmutl H C, Spink A, Ozmutlu S. Analysis of Large Data Logs: An Application of Poisson Sampling on Excite Web Queries[J]. Information Processing & Management, 2002, 38(4): 473-490.

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: zhangxiaojuan624@gmail.com。

[1] 张晓娟. AOL.zip. AOL 查询日志数据.

[2] 张晓娟. Labeled data.sql. 抽样与标注后的信息类、导航类与事务类查询.

[3] 张晓娟. results.xls. 三类别查询的不同网络动态性分析结果.

收稿日期: 2016-11-07
收修改稿日期: 2017-02-13

Analyzing Dynamic Informational, Navigational and Transactional Online Queries

Zhang Xiaoojuan
(School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: [**Objective**] This paper aims to improve the performance of search engines optimization through analyzing dynamic informational, navigational and transactional online queries. [**Methods**] First, the author analyzed user intentions with queries, Web documents and the information needs. Second, for each category of query intention, this paper investigated the changing of Web documents and information needs for different trending queries. [**Results**] The distribution of popular informational, transactional and navigational queries were different. The informational queries were more dependent on Web documents and needs than the other two types of queries. [**Limitations**] The data for this study was collected in 29 days. More research is needed to automatically identify and aggregate the popular queries. [**Conclusions**] Search engines need to list diversified results for informational queries. They need to keep the relevant pages on the first page for navigational queries, maintain the original ranking of relevant pages for the user behavior-related queries, and improve the novelty of results for the entertainment-related queries.

Keywords: Informational Query Transactional Query Navigational Query Query Dynamic Information Need Dynamic Document Content Dynamic